Iranian Traditional Music Subgenre (Dastgah) Recognition Using Ensemble Learning And Graph-Based Representation By Introducing New Database

Sina Ghazanfaripour[a], Morteza Khademi[a]*, Abbas Ebrahimi-Moghadam[a]

* Corresponding author, Email: khademi@um.ac.ir, Tel: +989153156497

ghazanfaripour@mail.um.ac.ir (S.Ghazanfaripour)

a.ebrahimi@um.ac.ir (A.Ebrahimi-Moghadam)

[a] Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

Abstract

Music plays a major role in daily life and serves as a key means for expressing human emotions. Automatic classification of Iranian traditional music is a fascinating yet challenging subject, particularly for those interested in Iranian music dastgahs. This paper proposes a novel method for Iranian traditional music genre recognition using Persian music tracks. Six Iranian music genres, namely Shour, Nava, Mahour, Segah, Chahargah, and Homayoun, are considered. To accurately detect genres, convolutional neural networks (CNNs), one-dimensional convolutional neural networks (1DCNNs), and long short-term memories (LSTMs) are employed. All models are fed extracted pitch features, with the music pitch converted into a sequential note vector and a visual representation in the form of a graph illustrating the musical structure. Finally, an ensemble model combines the predictions from all models. The proposed approach is evaluated using the "Arg" database, which includes solo melodic instrument tracks with no limitations on playing style, instrument, tempo, or techniques. The proposed method achieved a recognition accuracy of 77.35%, which improved to 80.44% with the use of data augmentation techniques. The experimental results, including accuracy, F1-score, and standard deviation (STD), demonstrate the effectiveness of the approach, showing better performance compared to other methods for dastgah recognition.

Keywords Dastgah Recognition, Deep Learning, Note Sequence Vector, Pitch, Graph, Data Augmentation

1 Introduction

Music significantly influences human emotions and mental states, prompting extensive research in areas like genre, emotion, instrument recognition, and music retrieval. Genre recognition, a complex pattern recognition task, has gained importance with growing music databases. Since genres overlap structurally and content-wise, especially in Iranian music,

labeling becomes difficult. Classifying tracks based on genre, content (such as instrumentation or solo/ polyphonic music tracks) and playing techniques is a key task in music signal processing. This paper focuses on recognizing Iranian music genres.

Iranian traditional music encompasses subgenres defined by structure, style, instrumentation, and regional roots such as dastgahs, avazes, religious, radif, and folk music. Dastgahs are subgenres based on unique note combinations and musical structures that set them apart from Western styles. Their complexity, coupled with limited pattern recognition studies in music, makes this area compelling. As traditional music is part of human heritage, access and awareness are vital. Yet, Iranian genre recognition remains underexplored due to its complexity and limited public familiarity. Identifying dastgahs has practical uses from recommendation systems (e.g., Shazam) and automatic classification to music education and cultural promotion. Recognizing dastgahs requires more than music theory; only experienced individuals with deep knowledge of their structures can identify them.

There is little research on Iranian traditional dastgah music compared to Western music, which benefits from its simpler structure and greater popularity. In western music, researchers have used different features like rhythm and tempo [1] to classify music tracks by genre, such as pop, classical, jazz, hip-hop and metal. In Iranian music, using such features for dastgah detection usually does not lead to accurate results due to different scales, the presence of quartertones, and varied rhythms, even within the same dastgah. Other features like Shahed note (a prominent note heard more frequently) can be used in Iranian music recognition.

Expert-based classification of Iranian dastgah music is time-consuming and difficult, highlighting the need for automated analysis to extract features and categorize dastgahs for easier music access. A fuzzy similarity study [2] shows that similarities between dastgahs range from 43% to 73%. Traditional automatic genre classification extracts key features from music signals, then applies a classification method. Recent techniques include Support Vector Machines (SVM), Neural Networks (NN), and Recurrent Neural Networks (RNN). Deep learning, which has advanced rapidly due to large datasets and powerful GPUs, offers the advantage of integrating feature extraction and classification within a unified structure.

In this paper, the proposed solution for identification of Iranian musical dastgahs includes two stages: note sequence vector formation and classification. Pre-processing steps such as windowing and filtering are required to form the note sequence vector. Windowing is used to prevent changes in the dynamics of music signal so that it can be regarded as stationary.

Filtering removes noise from the music signal and increases Signal-to-Noise Ratio (SNR). YAAPT (Yet another algorithm for pitch tracking) is employed for pitch detection, extracting the fundamental frequencies. This information is then used to create a note sequence vector and a visual representation of the audio signal, which serve as inputs to the classifiers. A multimodal classifier combining convolutional neural network (CNN), one-dimensional convolutional neural network (1DCNN), and long short-term memories (LSTM) in parallel is used to analyze various aspects of the music signal, such as temporal dependencies and the visual representation. Classification accuracy is further improved with an ensemble model using the ARG database. The details of this new database are fully explained in Section 4.1. The designed system classifies tracks into six dastgahs of Mahour, Shoor, Nava, Segah, Homayoun, and Chahargah. To the best of our knowledge, this is the first research that has been done in recognizing Iranian musical dastgahs with regard to the internal subtleties of this type of sub-genre with different compositional diversity by an ensemble method and data augmentation.

The paper is structured as follows: Section two presents related works, section three details the proposed architecture and methodology, section four includes simulation results, and section five offers conclusions and future work.

2 Related Work

Many researchers have studied western music genre classification, employing machine learning for effective solutions in tasks like beat detection and emotion and chord recognition [3-5]. The classification of western music genres was introduced as a recognition issue in 2002, suggesting features such as timbre, rhythm, and pitch [6]. Hand-crafted features were designed to extract specific attributes, with various classifiers used for genre identification [7]. SVM classification and Mel spectral coefficients were employed for classifying genres like Chinese folk, rock, pop, and Guzheng, with optimization attempted via a cuckoo search algorithm [8]. Other methods included Relevance Vector Machine (RVM), decision tree, and K Nearest Neighbor (KNN). Spectral features of classical, folk, Ghazal, and Sufi genres were mapped to an emotional plane for classification [9]. Conditional random fields modeled Chinese folk music, while Gaussian Mixture Models (GMM) helped identify track genres [10]. Features like Spectral Rolloff (SR), Zero Crossings (ZC), pitch, and duration also assisted in genre identification [11]. The KNN, Naive Bayes, neural network, and SVM methods were employed for classifying the GTZAN and Free Music Archive (FMA) databases. A hybrid method was

also utilized to address high genre overlap, which caused incorrect music track identifications [12].

Deep learning is gaining attention lately because powerful computing systems are now more widely available. Most deep-learning methods automatically extract distinctive information from data samples, unlike handcrafted features. The first use of CNN in music information retrieval occurred in 2012 [13]. Subsequent applications included automatic detection of music chords [14], onset detection [15], music classification based on spectrogram characteristics [16], and addressing challenges in music information retrieval [13]. In [17], harmonic and percussive components of music signals were separated before spectrogram generation and CNN training. CNN training can be time-consuming, prompting researchers to focus on optimizing network parameters [18]. The combination of handcrafted features and CNN properties was explored for music genre distinction [19], as well as combining handcrafted image features (spectrograms) with audio signal features for genre classification [20].

Generalized deep learning methods, including those with statistical structures that can conditionally enable or disable specific parts of the network, have found applications in music analysis [21]. A modified CNN version has been adapted for classifying music across various databases, including GTZAN and FMA [22-23]. CNNs can also be combined with RNNs [24]. For music onset detection, CNNs can automatically extract features from raw data with less manual preprocessing but higher computational costs, making them comparable to RNNs [15]. A residual network (ResNet) was used in [25], highlighting the challenges posed by high genre overlap in Iranian music dastgahs. The model assigns three genres based on statistical analysis of a 3-second track from the GTZAN database.

Shortening track durations in small databases increases training data and improves system performance [26]. Research explored the effect of varying track lengths on genre recognition, using five architectures for 1DCNN [27]. Nigerian databases with 1000 thirty-second tracks in seven genres were categorized using 1DCNN [28]. The FMA database was analyzed with 1DCNN, CNN, LSTM, and ensemble methods, utilizing two result-mixing techniques [29]. Four ensemble methods were considered, with a voting mechanism achieving the best results [30]. The nature of the XGBoost ensemble method was also discussed [31].

Since a piece of music is a sequence of notes played by a particular rule, LSTM and Gated Recurrent Unit (GRU) can be used, which are suitable for solving time-based signal processing problems [32]. In these studies, datasets including GTZAN, Emotify, Ballroom, and LastFM

were utilized, with Mel-spectrograms used for feature extraction. The integration of CNN and LSTM was also investigated [21].

A variety of optimizers can also be used to distinguish music genres. For example, in a research on GTZAN and Magna TagAtune databases, CNN and Self Adaptive Sea Lion Optimizer (SA-SLnO) were used for classification and system performance improvement, respectively [33]. In the mentioned methods, various types of libraries such as Librosa [16, 28, 30], Essentia [12], and Alize [34] were considered for extracting various features.

This section highlights key studies on Iranian music in the Music Information Retrieval (MIR) field, with fewer than 20 studies in the last 20 years [35]. The first research using an artificial neural network aimed to identify the Mahour dastgah from five others using an RBF network on FFT peaks [36]. Additionally, an RBF network and SVM classifier trained on pitch, MFCC, and Spectral Centroid were used for dastgah recognition [37]. The Nava dataset, MFCC features, and SVM classifier were applied for dastgah and instrument recognition, though with limited success [38]. GMM also analyzed 143 performances using spectrogram and Chroma features [39].

The first deep learning attempt for dastgah recognition used combination of CNN and GRU on STFT outputs, based on a dataset with 1137 solo violin and Ney tracks [40]. Another CNN model was trained on solo and ensemble performances across Iranian instruments [41]. A combined CNN and residual network was applied on the PMG dataset but limited by training examples [42]. Sequential note extraction, hierarchical classification, and LSTM networks classified music into 6 dastgahs and 11 sub-dastgahs, utilizing the ARG dataset of solo melodic instrument tracks [43].

To summarize, both deep learning and classical machine learning methods have been effectively applied to genre classification, with MFCCs and SVMs being prominent features and methods, especially in Indian music. The GTZAN database is the most used in literature. Research on Iranian music is scarce, often focusing on limited genre classes and overlooking the unique characteristics of Iranian dastgahs.

3 Proposed method

Fig. 1 shows the block diagram of the proposed method. The music track (S[n]) served as the input signal, undergoing a windowing stage to create L music segments ($S_i[n]$, i=1: L). Subsequently, pitch features ($F0_i$) were computed, forming a note sequence vector ($N_1, \ldots, N_L$),

and a graph-based model for each music track. Further, the outputs of 1DCNN, LSTM, and CNN models were employed to achieve better results through ensemble learning.

3.1 Windowing

Music signals are inherently non-stationary, with their frequency characteristics changing over time. To analyze such signals effectively, we divide them into shorter segments where they can be approximated as stationary. This is achieved through a process known as windowing, where overlapping windows are applied to capture musical nuances, such as notes, without losing important information.

Different window types have specific advantages and drawbacks. Rectangular windows cause edge discontinuities and distortions, so windows tapering toward zero at the ends and one in the center are preferred. Window length is as important as its type. it must be short enough to capture a single note. Shorter windows offer better time resolution, while longer ones improve frequency resolution. To avoid information loss, overlapping windows typically 25% to 75% are used [44]. More overlap ensures continuous signal coverage but increases computational complexity.

In this paper, Hanning window was utilized to produce music segments (Si[n]). Hanning window is a standard choice in signal processing due to its proven effectiveness. It is widely preferred for its ability to create smooth segments, minimize spectral leakage, and reduce edge effects. This leads to more accurate frequency representation, which is vital for distinguishing subtle pitch variations in non-stationary signals like music.

3.2 Pitch detection

Pitch is a core feature of music signals, reflecting the fundamental frequency (F0) and perceived note. Although F0 varies significantly and is useful for tasks like emotion, gender, and genre recognition, it is difficult to detect due to issues such as pitch doubling or loss at low frequencies, especially in poor recordings. Pitch directly maps to musical notes and is vital for capturing the melodic structure of Iranian dastgahs, making it well-suited for sequence-based learning models. In contrast, features like MFCC are less suitable due to the complexity of melodic patterns and increased computational complexity.

This study uses note sequences as classifier inputs, as they reflect the structural differences among dastgahs, unlike rhythm or timbre, which are more relevant to emotion [45–46] and instrument recognition [47], respectively. A key idea of this paper is to provide a meaningful

graph representation of note sequences, which further enriches this approach. Pitch detection methods fall into time-domain, frequency-domain, and hybrid types. Algorithms like YAAPT [48], SPICE (Self-supervised Pitch Estimation) [49], RAPT (Robust Algorithm for Pitch Tracking) [50], and YIN [51] were evaluated, with YAAPT chosen for its better accuracy. Fig. 2 illustrates YAAPT's steps, which combine time and frequency analyses to enhance F0 estimation, addressing limitations of methods like RAPT.

To clarify the pitch detection choices in this study, a bandpass filter was applied during preprocessing to retain frequencies between 26 and 2000 Hz (covering notes A0 to B6). This range includes 75 main musical notes, ensuring broad frequency coverage while removing irrelevant low/high-frequency noise. F0 candidates were identified using spectral harmonics correlation [52]. Other techniques like normalized low-frequency energy ratio [53] could also be used.

3.3 Note mapping

In the next step, the fundamental frequencies extracted from the music segments ($F0_i$) were converted into notes using a "note mapping" table, considering the sound range of various instruments across up to seven octaves. Table 1 illustrates the mapping of notes for fourth octave frequencies in Iranian music as an example. In this table, the terms Koron, Sori, and Bemol represent Quarter flat, Quarter sharp, and Flat, respectively.

This paper expands musical note extraction beyond western music's main seven primary notes and five semitones by incorporating semitone and quartertone intervals in Iranian traditional music. This inclusion results in finer frequency intervals, making note recognition in Iranian music more complex. While both the central frequency and surrounding band length vary for each note, the note mapping method primarily relies on central frequency values. Notes were assigned by calculating the minimum Euclidean distance between each $F0_i$ and adjacent notes for precise matching. Subsequently, a vector of length L was generated from the sequence of assigned notes across each music track and was then fed into the classifiers.

3.4 1DCNN-based classification

The main stages of CNNs are feature extraction and classification. 1DCNN is selected for its ability to efficiently handle sequential patterns in music signals, particularly for recognizing temporal patterns in music. Compared to CNNs used in parallel branches, 1DCNN reduces computational complexity, processes sequential note data directly, and achieves faster

performance while maintaining high accuracy. This is especially beneficial for recognizing complex melodic structures in Iranian musical dastgahs.

The proposed 1DCNN model for dastgah classification, shown in Fig. 3, was designed through trial and error based on other architectures [27]. It consists of 10 layers, plus input and output layers. The model includes two convolutional layers with 32 filters and a kernel size of three, followed by a max-pooling layer with a pool size of three. This is followed by two convolutional layers with 64 filters, another max-pooling layer with a pool size of five, and a final convolutional layer with 128 filters. There are three fully connected layers with 256, 128, and 64 neurons. Convolutional layers detect local patterns, and ReLU activation functions introduce non-linearity. Pooling layers reduce feature map size and complexity, making features more robust. The Softmax function is applied at the output layer with six neurons, corresponding to the six classes.

3.5 Note selection

In this step, based on the classifier type and computational costs, a smaller input size was constructed as a vector. By performing this operation, a new vector with R (where $R < L$) notes was created for each track, serving as an abstraction of the previous vector. To achieve this, the sequence vector containing L extracted notes is divided into equal note intervals. The interval size was determined based on the presence of at least one note every 0.5 seconds and having an odd number of elements. Additionally, by choosing the middle element of each interval as the representative note, the final vector size closely approximates the actual number of played notes in music tracks. It is worth mentioning that the last note in this vector is a repetition of the penultimate note. This vector serves as an input for the LSTM stage in Fig. 1.

3.6 LSTM-based classification

An LSTM network is a modified form of RNN, capable of learning long-term dependencies. It overcomes the issue of vanishing/exploding gradients by parametrizing the LSTM network using weight matrices from the input and previous state for each gate along with the memory cell [54]. In LSTMs, we use a logistic function as a gate function and a hyperbolic tangent as an activation function. In this paper, LSTM is chosen for its ability in learning relationships of notes sequences, long-range dependencies and unraveling complex patterns in the music signal. Compared to GRUs, which are efficient but less effective at managing long-term dependencies, LSTMs are more suitable for handling intricate musical patterns due to their superior capability to maintain and utilize long-term memory.

The block diagram of the employed LSTM architecture in this work is shown in Fig. 4. The input is note sequence vector with length of R. The model has four LSTM layers with dropouts in between. Dropouts are applied to mitigate overfitting. After that, the output is passed to the fully connected network. SoftMax is used to classify the final output in each stage by assigning probabilities to each class label based on the learned features from the LSTM layers. The first two LSTM layers have 128 units, the third has 64 units, and the fourth has 32 units. The first two layers extract basic features from the note sequences, while the next two layers extract higher-level features and long-term relationships within the sequence.

Due to the overlap of Iranian music dastgahs and based on simulation results, a hierarchical model was proposed [43]. This model is illustrated in Fig. 5. In the first stage, LSTM1 recognizes the Mahour dastgah. In the second stage, LSTM2A recognizes Shur and Nava, while LSTM2B classifies Homayoun. In the third stage, LSTM3 identifies Segah and Chahrgah. The classification path is determined based on each classifier's decision. For example, a track from Segah enters LSTM1, which classifies it into one of the three classes: Mahour, Shur/Nava, or Homayoun/Segah/Chahrgah. If correctly classified, it proceeds to LSTM2B, which determines if the track is Homayoun or Segah/Chahrgah. Finally, LSTM3 classifies the track, and the process ends for that track. The next track follows the same process. All classifiers are trained using labeled data specific to their task.

3.7 Graph representation

Fig. 6 shows the visual representation of music, where the note sequence vector is displayed in a circular format. This circular graph, based on musical octaves, is the main contribution of this paper. It includes three elements: nodes (musical notes), links (order of notes), and node colors (indicating note repetition rate). The circular representation was chosen because it allows equal positioning of nodes and makes it easier to show links between related nodes. For a square form, the nodes positioned on the edges have different placements compared to those on the sides. Additionally, as the number of links increases, displaying connections between nodes on the same side becomes complicated. This issue does not arise in a circular representation, where all nodes are placed equally and link display is simpler.

Each music track generates a unique graph pattern within a circular structure, differing in link placement and note repetition. This representation helps identify the Shahed note "prominent note heard more than other notes in a music track" and serves as an intuitive alternative to the spectrogram, displaying conceptual and comprehensible features like played

notes, sequence of notes and Shahed note. It can also be utilized for music track classification with classifiers such as CNN. Fig. 7 illustrates three examples, highlighting differences in node colors, link numbers, and connections, demonstrating the distinct patterns of each dastgah.

3.8 CNN-based classification

Artificial neural networks (ANNs) consist of neurons that learn and adapt, forming the basis of CNNs. CNN excel at processing large sequential data, such as images represented by pixel values [55-56]. They perform convolutions in a superposed space using smaller kernels, and padding with zeros allows for consistent feature map sizes across multiple convolutions. Techniques like Max Pooling help reduce computational complexity while preserving essential information. As feature maps traverse various convolutional layers, the kernels learn to identify patterns and abstract features. Additionally, music signals can be transformed into image representations for music genre classification using CNN, which is favored for its robust feature extraction capabilities, essential for recognizing complex patterns in graph representations of music [17-20].

Fig. 8 illustrates the CNN architecture used in the proposed approach, consisting of four convolutional (Conv) blocks with 64, 128, 256, and 512 filters. The Conv filters have kernel sizes of 5×5 and 3×3, with padding applied to all layers, followed by max pooling layers operating over 2×2 sub-windows. At the end of the network, six fully connected dense layers lead into the c-way Softmax classification layer. When back-propagation occurs, the maximum amount of error is passed under the Relu activation due to its unity gradient. The model's performance was enhanced through dropout regularization after each layer. This architecture was refined through empirical trial and error to achieve an optimal balance between feature extraction, model complexity, and performance in dastgah recognition. Thereafter, the CNN model results were fed to the proposed ensemble model for dastgah recognition.

3.9 Ensemble learning

This part of the proposed system uses ensemble learning by combining the outputs of base models (C1, C2, and C3) to determine the final class of each track. Ensemble learning improves accuracy by merging predictions from multiple models. It is categorized as homogeneous (same models) or heterogeneous (different models). The proposed system uses heterogeneous ensemble learning because the base models are built with different algorithms. Various techniques, including voting, averaging, boosting, bagging, and stacking, are used in ensemble learning based on prediction characteristics. This paper was used the stacking technique for

three reasons: it works well with heterogeneous ensemble models [57], helps reduce overfitting and enhances model robustness, and leverages the strengths of each base model by combining diverse outputs. This diversity enables the ensemble to handle different scenarios effectively.

Unlike other ensemble techniques, stacking involves a meta model, which can be any machine learning model, such as a support vector machine, neural network, or linear regression. Meta model learns how to combine the base models' predictions. In this study, base models (C1, C2, and C3) provide their output probabilities for each class as inputs to meta model, allowing it to learn a better combination of these outputs based on the training data to enhance performance metrics like F1 score, accuracy, and standard deviation (STD). During training phase, the meta model learns the relationships between C1, C2, C3, and the actual labels and optimizes parameters with algorithms like gradient descent and Adam to minimize loss functions such as cross entropy.

In this research, the meta model comprises three fully connected layers and a Softmax layer. The fully connected layers extract features from the input, while the Softmax layer performs statistical analysis for each class, ultimately deciding on the dastgah classification.

4 Experimental results

The proposed approach was evaluated on the Arg database, which consists of 20-second tracks (S(n)) as input signals. The input size for 1DCNN and LSTM was 2000 (L) and 45 (R) notes vector respectively. Different notes were extracted from frequencies ranging from 26 Hz to 2000 Hz using a third order Butterworth bandpass filter. This filter was chosen for its smooth and flat response, which preserves signal quality. Its simple design facilitates implementation, and it effectively passes the desired frequency range while attenuating frequencies outside this range. We have quantified the performance of 1DCNN, CNN, and LSTM models individually. A performance estimate has also been obtained for the ensemble model. A k-fold cross validation test [58] with $k = 5$ was used to assess the performance of various models. A batch size of 32, chosen for its balance between efficiency and memory usage and ensuring no remaining data, was used for every model, which was trained for 50 epochs. For training the proposed models, a learning rate of 0.001 was selected. The Cross-Entropy loss function and the Adam optimizer were then employed to optimize the model parameters and achieve efficient convergence during the classification task.

The proposed method was implemented with TensorFlow and ran on a PC with Intel(R) Core(TM) i7-4500U CPU @2.40 GHz, 16GB memory, NVIDIA GeForce GTX and 64-bit Windows 10 operating system.

4.1 Arg database

In Western music, databases such as the Latin Music Database, Ballroom, and ISMIR2004 exist, with GTZAN being the most popular for genre recognition, containing 1,000 music tracks across 10 genres. However, Iranian music lacks a standardized database with consistent track lengths, distribution, and labels. The only existing database for Iranian music contains 1143 tracks, each 30 seconds long, and focuses exclusively on the Mahour dastgah [59]. To address this gap, this research developed the "Arg" database for Iranian dastgah recognition. According to our surveyed musicians and composers, listeners familiar with the music can identify a dastgah within 20 seconds. Consequently, the Arg database consists of 20-second tracks, primarily featuring solo instrument recordings played by professional musicians (e.g., piano, tar, ney, oud, santour, and violin), which minimizes interference from other instruments and eliminates the need for source separation. No percussion solos were included.

This research limited its study to six dastgahs: Chahargah, Segah, Homayoun, Shoor, Nava, and Mahour, due to the inherent complexity and overlap in Iranian classical music's seven dastgahs. Because Mahour and Rast-Panjgah are difficult to distinguish due to similar scales, both were labeled as Mahour in the Arg database. All dastgahs in the Arg database encompass various goushehs. The database comprises 900 20-second tracks, accurately labeled by composers and validated by Iranian traditional music experts, offering potential for competitive use in research. However, limitations exist, including data quantity, varied data sources, and incomplete metadata on recording conditions and equipment. The distribution of music tracks includes 150 tracks for each of the six dastgahs in this database.

To augment data in audio signal processing, various techniques can be employed to increase dataset diversity, improve model generalization, and enhance the performance of machine learning models. One fundamental method involves multiplying the audio signal by a random amplitude factor, such as adjusting the volume between -12 dB and 12 dB to introduce variability. Adding noise, like white noise at levels of 0.01 and 0.05 and Gaussian noise between 0.005 and 0.02, also improves training data diversity and model robustness. Additionally, reverberation effects enhance audio textures, further aiding model generalization.

These techniques allow each audio segment to be transformed into multiple versions, increasing the database size and improving performance in predicting musical patterns.

By employing these augmentation techniques, each audio segment can be transformed into five new versions by increasing and decreasing the volume, adding white and Gaussian noise, and adding reverberation, thereby increasing the total number of tracks from 900 to 4500 (900 original + 3600 new). First, the tracks is split into 720 tracks for training and 180 for testing using k-fold cross-validation (k=5). In each step, data augmentation is applied exclusively to the training set. Consequently, the training set comprises 720 original tracks and 3600 new tracks. Therefore, there are 4320 tracks for training and 180 tracks for testing (only original tracks without augmentation). The average of these 5-fold results shows the overall performance of the proposed method.

In summary, employing these augmentation strategies enhances database diversity, improves the model's generalization capabilities, prevents overfitting, and leads to better performance in pattern recognition tasks.

4.2 Evaluation results

Several parameters were evaluated to determine the performance of individual models, including accuracy, F1 score and standard deviation. For a multi-class classification problem, the F1 score is adapted using methods like macro and micro averaging. In this paper, macro-averaging was used due to the balanced database, where each class has a similar number of instances. Model stability is indicated by the F1 score, as a stable model produces similar predictions for similar inputs. A low STD indicates repeatability and suggests consistent overall performance across evaluations.

Fig. 9 presents the classification accuracy for dastgah recognition across different window lengths, types, and overlaps. Based on this figure, Hanning window, 20-millisecond music segments (Si(n)) with a 50% overlap offers the highest classification accuracy. This values ensure adequate time and frequency resolution, accurately extracting pitches and analyzing musical content.

In terms of pitch detection accuracy, as shown in table 2, YAAPT demonstrates better performance with an accuracy of 77.35%. This advantage is attributed to its strong noise resistance and precise feature extraction, making it well-suited for detecting complex musical notes in Iranian music. The SPICE algorithm follows closely with an accuracy of 76.91%,

though it is slightly less effective due to its sensitivity to noise and limited adaptability to the characteristics of Iranian notes. RAPT achieves an accuracy of 76.66%, hindered by issues with frequency resolution, while YIN, with an accuracy of 75.83%, struggles with subtle pitch variations and exhibits higher error rates in noisy environments. Overall, YAAPT stands out as the best option for accurate pitch analysis in complex compositions due to its advanced techniques and adaptability.

The performance of these four pitch detection algorithms was evaluated for detecting quarter-tone notes in Iranian music by analyzing 12 specified Koron and Sori notes within the 330–650 Hz frequency range. Each note's precise frequency was taken as the reference, and the accuracy of frequency estimation by each algorithm was compared to these values. As illustrated in Figure 10, YAAPT stands out for detecting quarter-tone notes, which play a crucial role in Iranian music. SPICE and RAPT also demonstrated higher accuracy than YIN. These findings can contribute to enhancing pitch detection methods in Iranian music and improving the generalizability of current models.

Here, we focus on the LSTM results before discussing the all classifiers performance. A hierarchical structure was implemented for the LSTM due to limitations in managing many outputs in the last layer. A single-layer LSTM showed poor performance, achieving a detection accuracy of only 49.6% for classifying six dastgahs. Adding a layer with 256 neurons improved this to 51.4%, but results were still disappointing. The hierarchical design aimed to reduce output numbers in the last layer. However, except for LSTM1, error propagation in each class decreased accuracy in subsequent categories, with detailed results presented in Table 3.

The performance evaluation of the proposed method encompassed a detailed analysis of CNN, LSTM, and 1DCNN models for six dastgah recognition. Fig. 11 illustrates the accuracy and the loss over 50 epochs, measured on the validation dataset (120 tracks of training set). Notably, the CNN model achieved an accuracy of approximately 75.14% with a loss of 0.275, the LSTM model demonstrated an accuracy of 74.84% with a loss of 0.258, and the 1DCNN model attained an accuracy of 76.71% with a loss of 0.186. An accuracy and loss curves saturated with increasing epochs. These results collectively provided a comprehensive overview of the proposed method's efficacy in dastgah recognition and form a robust foundation for meaningful comparisons with other architectures.

Table 4 tabulates the performance metrics for individual and ensemble models in dastgah recognition on the ARG database. The 1DCNN (C1), LSTM (C2), and CNN (C3) models

showed individual accuracy, but the ensemble method significantly outperformed the base models. The stack-based ensemble approach achieved 77.35% accuracy, benefiting from the strengths of the base models. Stacking enhances prediction by combining complementary model strengths, with a low standard deviation indicating reliability. In experiments with five ensemble methods, XGBoost showed high accuracy and fast prediction times but longer training. AdaBoost and Bagging are simpler models with lower accuracy. Stacking offers high accuracy but requires more computation, while Majority Voting is the fastest but less accurate.

Table 5 shows the experimental results on the ARG database with data augmentation, demonstrating an improvement in classification accuracy for all models. The stack-based ensemble learning model achieved 80.44% accuracy, an increase of about 3%. This improvement is due to data augmentation techniques that expanded the training data, allowing the models to learn more comprehensive patterns and generalize better. An 80.44% accuracy in recognizing Iranian musical dastgahs is valuable for music education, archiving, classifying tracks, musicological research, identifying dastgahs in performances, and interactive educational tools. This accuracy is practical for many educational and research purposes.

The comparison between the proposed graph-based method and the Mel-Spectrogram was performed using a proposed CNN model for both approaches. For the Mel-Spectrogram, 25ms overlapping windows with a 10ms overlap and 128 Mel filters were used, resulting in a 128x128 spectrogram. The classification accuracy for six Iranian musical dastgahs showed that the graph-based method achieved 72.54%, improving to 76.56% with data augmentation. In contrast, the Mel-Spectrogram reached 68.03%, improving to 70.92% with data augmentation. These results indicate that the proposed graph model can be an effective tool for analyzing Iranian music.

The average confusion matrix of the LSTM, 1DCNN, CNN and ensemble learning for detecting six dastgahs is shown in Fig. 12. Although the performance of the proposed system is notable for some dastgahs such as Mahour, but it was difficult to categorize some similar dastgahs. For example, the accuracy of the system in detecting the Nava was the lowest, and the relatively large numbers in the crossing cells of Shour and Nava showed this classification difficulty due to the similarity of the dastgahs. This issue had a negative effect on the average accuracy of the system. To measure this negative effect, we integrated two dastgahs and test a 5-class classification by the system. Under these conditions, the average accuracy of proposed method was improved by about 6% and reached from 80.44% to 86.67%.

Table 6 presents a comparison of our experimental results with previous genre detection methods, highlighting the feature extraction method, classification type, music region of interest, and the publication year. This comparison has been done entirely on the Arg database. The proposed method performed better than previous approaches, primarily due to its hybrid approach, which simultaneously addresses the weaknesses of base model classifiers and focuses on semi-quarter tones in pitch detection. Additionally, methods for identifying Western music genres cannot provide an appropriate response to the issue of dastgah as a sub-genre of Iranian classical music. Moreover, some methods concentrate on the main genres of Iranian music, such as traditional, folk, pop, and rock, which exhibit much less overlap compared to the dastgahs.

The performance of the proposed model is not affected by changes in tempo, instrument, or playing style. This is because the extracted pitch features from the music signals are specifically chosen for identifying musical dastgahs and are not sensitive to these variations. Furthermore, the database containing tracks with diverse tempos, styles, and instruments, the model evaluations show that the model can effectively recognize dastgahs under various and diverse conditions. It is worth noting that other results are presented in the appendix.

5 Conclusion

This paper proposes the ensemble learning-based dastgah recognition system. Firstly, the fundamental frequency was extracted from music signals. After that, these features were fed to 1DCNN, CNN and LSTM based models. The ensemble model was proposed based on the CNN, LSTM, and 1DCNN output. The experimental results suggested that the ensemble model achieves better classification performance than the other models employed in the proposed approach. The proposed ensemble model outperformed the compared methodologies with 77.35% accuracy for dastgah recognition on the Arg database. The classification accuracy increased to 80.44% with data augmentation. The proposed method was also evaluated on GTZAN database and obtains 73.29% accuracy using the same features and model parameters. All the models provided impressive accuracy individually and showed a low STD. The experimental results suggested that the proposed approach is suitable for this purpose and paves the way for upcoming research fields such as music retrieval, instrument classification, transcription and emotion recognition.

Availability of data

The dataset generated is available in https://zenodo.org/doi/10.5281/zenodo.12648387

Declarations

Conflict of interests: The authors declare that they have no conflicts of interest.

The Supplementary data is available at:

file:///C:/Users/pc/Downloads/Khademi-25-SCI-2410-9529-%20supplementary%20file-3.pdf

References

[1] Mayer, R., Neumayer, R., and Rauber, A., "Combination of audio and lyrics features for genre classification in digital audio collections." in Proceedings of the 16th ACM international conference on Multimedia, New York, NY, USA, pp. 159-168, (2008). https://doi.org/10.1145/1459359.1459382

[2] Abdoli, S., "Iranian Traditional Music Dastgah Classification.", in Proceedings of the 12th International Society for Music Information Retrieval Conference, Miami, United States, pp. 275–280, (2011). https://doi.org/10.5281/zenodo.1417425

[3] Birajdar, G. K., and Patil, M. D., "Speech/music classification using visual and spectral chromagram features. Journal of Ambient Intelligence and Humanized Computing." **11**(1), pp. 329-347. (2020). https://doi.org/10.1007/s12652-019-01303-4

[4] Foleiss, J. H., and Tavares, T. F., "Texture selection for automatic music genre classification. Applied Soft Computing." 106127. **89**, (2020). https://doi.org/10.1016/j.asoc.2020.106127

[5] Chang, W.-H., Li, J. L., Lin, Y. S., and Lee, C.C., "A GENRE-AFFECT RELATIONSHIP NETWORK WITH TASK-SPECIFIC UNCERTAINTY WEIGHTING FOR RECOGNIZING INDUCED EMOTION IN MUSIC." 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE, San Diego, CA, USA, pp. 1-6, (2018). https://doi.org/10.1109/ICME.2018.8486570

[6] Tzanetakis, G., and Cook, P., "Musical genre classification of audio signals." IEEE Transactions on speech and audio processing, **10**(5), pp. 293-302, (2002). https://doi.org/10.1109/TSA.2002.800560

[7] Bengio, Y., Courville, A., and Vincent, P., "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence, **35**(8), pp. 1798-1828, (2013). https://doi.org/10.1109/TPAMI.2013.50

[8] Shi, W. and Fan, X., "Speech classification based on cuckoo algorithm and support vector machines." 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), IEEE, Beijing, China, pp. 98-102, (2017). https://doi.org/10.1109/CIAPP.2017.8167188

[9] Chaudhary, D., Singh, N. P., and Singh, S., "Genre Based Classification of Hindi Music." International Conference on Innovations in Bio-Inspired Computing and Applications. IBICA 2018. Advances in Intelligent Systems and Computing, Springer, Cham, **939**. pp 73–82, (2018). https://doi.org/10.1007/978-3-030-16681-6_8

[10] Li, J., Ding, J., and Yang, X., "The regional style classification of Chinese folk songs based on GMM-CRF model." Proceedings of the 9th International Conference on Computer and Automation Engineering. Association for Computing Machinery, New York, NY, USA, pp. 66–72. (2017). https://doi.org/10.1145/3057039.3057069

[11] Kaur, C. and Kumar, R., "Study and analysis of feature-based automatic music genre classification using Gaussian mixture model." 2017 International Conference on Inventive Computing and Informatics (ICICI), IEEE. India, pp. 465-468, (2017). https://doi.org/10.1109/ICICI.2017.8365395

[12] Karunakaran, N. and Arya, A., "A Scalable Hybrid Classifier for Music Genre Classification using Machine Learning Concepts and Spark." 2018 International Conference on Intelligent Autonomous Systems, IEEE, Singapore, pp. 128-135, (2018). https://doi.org/10.1109/ICoIAS.2018.8494161

[13] Humphrey, E. J. and Bello, J. P., "Rethinking automatic chord recognition with convolutional neural networks." 11th International Conference on Machine Learning and Applications, Boca Raton, FL, USA, pp. 357-362, (2012). https://doi.org/10.1109/ICMLA.2012.220

[14] Humphrey, E. J., Bello, J. P., and LeCun, Y., "Moving beyond feature design: Deep architectures and automatic feature learning in music informatics." In Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, pp. 403-408. (2012). https://archives.ismir.net/ismir2012/paper/000403.pdf

[15] Schlüter, J. and Böck, S., "Improved Musical onset detection with convolutional neural networks. Improved Musical Onset Detection with Convolutional Neural Networks." In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, pp. 6979–6983, (2014). http://doi.org/10.1109/ICASSP.2014.6854953

[16] Athulya, K M., and Sindhu, S., "Deep learning based music genre classification using spectrogram." 2nd International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS 2021). (2021). https://dx.doi.org/10.2139/ssrn.3883911

[17] Gwardys, G. and Grzywczak, D. M., "Deep image features in music information retrieval." International Journal of Electronics and Telecommunications, **60**(4), pp. 321-326. (2014). http://doi.org/10.2478/eletel-2014-0042

[18] Sigtia, S. and Dixon, S., "Improved music feature learning with deep neural networks." 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). Florence, Italy, pp. 6959-6968, (2014). https://doi.org/10.1109/ICASSP.2014.6854949

[19] Costa, Y. M. G., Oliveira, L. S., and Silla Jr, C. N., "An evaluation of convolutional neural networks for music classification using spectrograms." Applied soft computing, **52**, pp. 28-38. (2017). https://doi.org/10.1016/j.asoc.2016.12.024

[20] Nanni, L. et al., "Combining visual and acoustic features for music genre classification." Expert Systems with Applications, Vol. 45, pp. 108-117. (2016). https://doi.org/10.1016/j.eswa.2015.09.018

[21] Chen, N. and Wang, S., "High-Level Music Descriptor Extraction Algorithm Based on Combination of Multi-Channel CNNs and LSTM." Proceedings of the 18th International Society for Music Information Retrieval Conference, Suzhou, China, pp. 509–514, (2017). https://doi.org/10.5281/zenodo.1417901

[22] El Achkar, C., Couturier, R., Atéchian, T., and Makhoul, A., "Combining Reduction and Dense Blocks for Music Genre Classification." Springer Nature Switzerland AG 2021 T. Mantoro et al. (Eds.): Neural Information Processing. ICONIP 2021. Communications in Computer and Information Science, Springer, Cham, **1517**, pp. 752–760, (2021). https://doi.org/10.1007/978-3-030-92310-5_87

[23] Liu, C., Feng, L., Liu, G., Wang, H., and Liu, S., "Bottom-up broadcast neural network for music genre classification." Multimedia Tools and Applications, Springer Science Business Media, LLC, part of Springer Nature. **80**, pp. 7313–7331 (2020). https://doi.org/10.1007/s11042-020-09643-6

[24] Feng, L., Liu, S., and Yao, J., "Music genre classification with paralleling recurrent convolutional neural network." arXiv preprint arXiv: 1712.08370. (2017). https://doi.org/10.48550/arXiv.1712.08370

[25] Bisharad, D. and Laskar, R. H., "Music Genre Recognition Using Residual Neural Networks." 2019 IEEE Region 10 Conference (TENCON 2019), Kochi, India, pp. 2063-2068, (2019). https://doi.org/10.1109/TENCON.2019.8929406

[26] Ndou, N., Ajoodha, R., and Jadhav, A., "Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches." 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, pp. 1-6, (2021). https://doi.org/10.1109/IEMTRONICS52119.2021.9422487

[27] Allamy, S. and Koerich, A. L., "1D CNN Architectures for Music Genre Classification." 2021 IEEE Symposium Series on Computational Intelligence (SSCI), Orlando, FL, USA, pp. 01-07, (2021). https://doi.org/10.48550/arXiv.2105.07302

[28] Falola, P. B. and Akinola, S. O., "Music Genre Classification Using 1D Convolution Neural Network." International Journal of Human Computing Studies, **3**(6), pp. 3-21. (2021). https://journals.researchparks.org/index.php/IJHCS/article/view/2108

[29] Kostrzewa, D., Kaminski, P., and Brzeski, R., "Music Genre Classification: Looking for the Perfect Network." Springer Nature Switzerland AG 2021 M. Paszynski et al. (Eds.): Computational Science – ICCS 2021, LNCS 12742, **12742**. pp. 55–67, (2021). https://doi.org/10.1007/978-3-030-77961-0_6

[30] Gupa, R., Yadav, J., and Kapoor, C., "Music Information Retrieval and Intelligent Genre Classification." Proceedings of International Conference on Intelligent Computing, Information and Control Systems. Advances in Intelligent Systems and Computing, Springer, Singapore, **1272**, pp 207–224, (2020). https://doi.org/10.1007/978-981-15-8443-5_17

[31] Islam, S., et al. "Machine Learning-Based Music Genre Classification with Pre-Processed Feature Analysis." Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI), **7**(3), pp. 491-502, (2021). http://dx.doi.org/10.26555/jiteki.v7i3.22327

[32] Ashraf, M., Geng, G., Wang, X., Ahmad, F., and Abid, F., "A Globally Regularized Joint Neural Architecture for Music Classification." IEEE Access, **8**, pp. 220980-220989, (2020). https://doi.org/10.1109/ACCESS.2020.3043142

[33] Kumaraswamy, B. B. and Poonacha, P. G., "Deep Convolutional Neural Network for musical genre classification via new Self Adaptive Sea Lion Optimization." Applied Soft Computing, 107446, **108**, (2021). https://doi.org/10.1016/j.asoc.2021.107446

[34] Rajan, R., et al. "Music Genre Classification Using Timbral Feature Fusion on i-vector Framework." INFOCOMP Journal of Computer Science, **20**(2), (2021). https://infocomp.dcc.ufla.br/index.php/infocomp/article/view/1604

[35] Ebrat, D. and Didehvar, F., "Iranian Modal Music (Dastgah) detection using deep neural networks." arXiv. (2022). https://doi.org/10.48550/arXiv.2203.15335

[36] Mahmoodan, S. and Banooshi, A., "Automatic classification of Mahoor scale (Dastgah) using artificial neural." 2nd Conf. Iran. Soc. Acoustic. Vib. (ISAV). Sharif Univ. Tehran. (2012). https://civilica.com/doc/188653 (in Persian)

[37] Abbasilayegh, M., Haghipour, S., and Sarem, Y. N., "Classification of the Radif of Mirza Abdollah a canonic repertoire of Persian music using SVM method," GAZI Univ. J. Sci. Part A Eng. Innov., **1**(4), pp. 57–66. (2013). https://dergipark.org.tr/tr/download/article-file/83761

[38] Baba Ali, B., Gorgan Mohammadi, A., and Faraji Dizaji, A., "Nava: A Persian Traditional Music Database for the Dastgah and Instrument Recognition Tasks." Journal of Advanced Signal Processing, **3**(2). pp. 125-134 (2019). https://doi.org/10.22034/jasp.2019.10444

[39] Heydarian, P. and Bainbridge, D., "Dastgàh recognition in Iranian music: Different features and optimized parameters." In Proceedings of the 6th International Conference on Digital Libraries for Musicology (DLfM '19). Association for Computing Machinery, New York, NY, USA, pp. 53–57, (2019). https://doi.org/10.1145/3358664.3361873

[40] Azar, S. R., Ahmadi, A., Malekzadeh, S., and Samami, M., "Instrument-independent Dastgah recognition of Iranian classical music using AzarNet." arXiv, pp. 1–5, (2018). https://doi.org/10.48550/arXiv.1812.07017

[41] Vafaeian, A., Borna, K., Sajedi, H., Alimohammadi, D., and Sarai, P., "Proposed Method for Note Detection and Automatic Identification of the Melody Models (Gusheh) in Iranian Traditional Music with Micro Approach." Eng. Manag. Soft Comput., **4**(2), pp. 41–66, (2018). https://jemsc.qom.ac.ir/article_1268.html?lang=en

[42] Farajzadeh, N., Sadeghzadeh, N., and Hashemzadeh, M., "PMG-Net: Persian music genre classification using deep neural networks." Entertainment Computing, 100518. ISSN 1875-9521, **44**, (2023). https://doi.org/10.1016/j.entcom.2022.100518

[43] Ghazanfaripour, S., Khademi, M., and Ebrahimi Moghadam, A., "Iranian Dastgah Music Recognition Based On Notes Sequence Extraction And Use Of LSTM Networks." Journal of Electrical and Computer Engineering, Iran, **20**(2), pp. 155-163. (2022). https://rimag.ir/fa/Article/29195 (in Persian)

[44] Helmrich, C. R. and Edler, B., "Audio Coding Using Overlap and Kernel Adaptation." IEEE SIGNAL PROCESSING LETTERS, **23**(5), pp. 590-594, (2016). https://doi.org/10.1109/LSP.2016.2538324

[45] Kang, J. and Herremans, D., "Are we there yet? A brief survey of Music Emotion Prediction Datasets, Models and Outstanding Challenges." ArXiv, abs/2406.08809. (2024). https://doi.org/10.48550/arXiv.2406.08809

[46] Sujeesha, S. A., Mala, J. B., and Rajan, R., "Automatic music mood classification using multi-modal attention framework." Engineering Applications of Artificial Intelligence, 107355, ISSN 0952-1976, **128**, (2024). https://doi.org/10.1016/j.engappai.2023.107355

[47] Gst, G., Mastrika, A. V., and Radhitya, M. L. "Musical Instrument Classification Using Audio Features and Convolutional Neural Network." Journal of Applied Informatics and Computing, **8**(1), pp. 226-234. (2024). https://doi.org/10.30871/jaic.v8i1.8058

[48] Kasi, K. and Zahorian, S. A., "Yet another algorithm for pitch tracking." 2002 ieee international conference on acoustics, speech, and signal processing, **1**, pp. I-361-I-364, (2002). https://doi.org/10.1109/ICASSP.2002.5743729

[49] Gfeller, B., Frank, C., Roblek, D., Sharifi, M., Tagliasacchi, M., and Velimirović, M., "SPICE: Self-supervised pitch estimation." IEEE/ACM Transactions on Audio, Speech, and Language Processing, **28**, pp. 1118-1128. (2020). https://doi.org/10.1109/TASLP.2020.2982285

[50] Talkin, D. and Kleijn, W. B., "A robust algorithm for pitch tracking (RAPT)." Speech coding and synthesis, pp. 497-518. (1995). https://www.ee.columbia.edu/~dpwe/papers/Talkin95-rapt.pdf

[51] De Cheveigné, A. and Kawahara, H., "YIN, a fundamental frequency estimator for speech and music." The Journal of the Acoustical Society of America, **111**(4), pp. 1917-1930. (2002). https://doi.org/10.1121/1.1458024

[52] Xuejing, S., "Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio." Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, FL, USA. pp. I-333-I-336, (2002). http://dx.doi.org/10.1109/ICASSP.2002.5743722

[53] Zahorian, S. A. and Hu, H., "A spectral/temporal method for robust fundamental frequency tracking." The Journal of the Acoustical Society of America, **123**(6), pp. 4559-4571. (2008). https://doi.org/10.1121/1.2916590

[54] Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., and Schmidhuber, J., "Lstm: A search space odyssey." IEEE Transactions on Neural Networks and Learning Systems, **28**(10), pp. 2222-2232, (2017). https://doi.org/10.1109/TNNLS.2016.2582924

[55] Topaloglu, I., "Deep learning based convolutional neural network structured new image classification approach for eye disease identification." Scientia Iranica, **30**(5), pp. 1731-1742. (2022). https://doi.org/10.24200/sci.2022.58049.5537

[56] Khosravani Pour, L. and Farrokhi, A., "Language recognition by convolutional neural networks." Scientia Iranica, **30**(1), pp.116-123. (2023). https://doi.org/10.24200/sci.2022.59110.6064

[57] Tahir, M. A., Kittler, J., and Bouridane, A., "Multilabel classification using heterogeneous ensemble of multi-label classifiers." Pattern Recognition Letters, **33**(5), pp. 513-523. (2012). https://doi.org/10.1016/j.patrec.2011.10.019

[58] Rodriguez, J. D., Perez, A., and Lozano, J. A., "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation." IEEE Transactions on Pattern Analysis and Machine Intelligence, **32**(3), pp. 569-575, (2010). https://doi.org/10.1109/TPAMI.2009.187

[59] Ghazanfaripour, S., Nezamabadi-pour, H., and Rashedi, E., "Content-based Persian music retrieval by Feature integration and GSA algorithm." First conference on swarm intelligence and evolutionary computation, Iran. (2016). https://doi.org/10.5281/zenodo.14892879 (in Persian)

Biographies

Sina Ghazanfaripour received his BSc and MSc degrees in Electrical Engineering from the Bahonar university of Kerman, Iran, in 2013 and 2016 respectively. Since 2018, he has been working toward the PhD degree in Telecommunications Engineering at the Ferdowsi University of Mashhad, Iran. His research interests include the area of audio processing, music information retrieval, machine learning, and pattern recognition.

Morteza Khademi was born in Iran, in 1958. He received the B.Sc. and M.S. degrees in electrical engineering from the Isfahan University of Technology, Isfahan, Iran, in 1985 and 1987, respectively, and the Ph.D. degree in electrical engineering and video communications from the University of Wollongong, Wollongong, NSW, Australia, in 1995. In 1987, he joined the Ferdowsi University of Mashhad, Mashhad, Iran. He is currently a Professor with the Department of Electrical Engineering, Ferdowsi University of Mashhad. His current research interests include video communications, biomedical signal processing, and data analysis.

Abbas Ebrahimi-Moghadam received his B.Sc. degree in Electrical Engineering from Sharif University of Technology, Tehran, Iran, and his M.Sc. degree in Electrical Engineering from K. N. Toosi University of Technology, Tehran, Iran in 1991 and 1995, respectively. He then received a Ph.D. degree in Electrical and Computer Engineering form McMaster University, ON, Canada. He is currently with Electrical Engineering Department at Ferdowsi University of Mashhad, Mashhad, Iran, as an assistant professor since 2011.

Figure and table captions list

Figures and tables

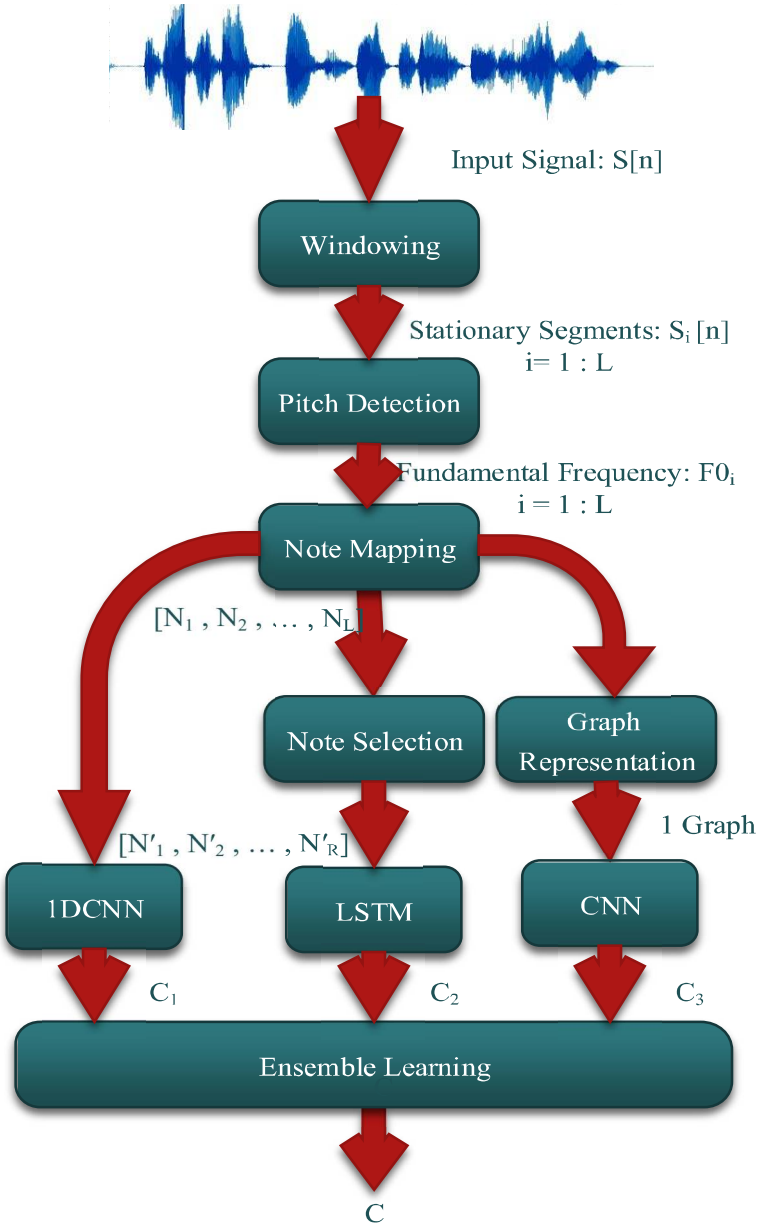

Fig. 1 Proposed method for dastgah recognition

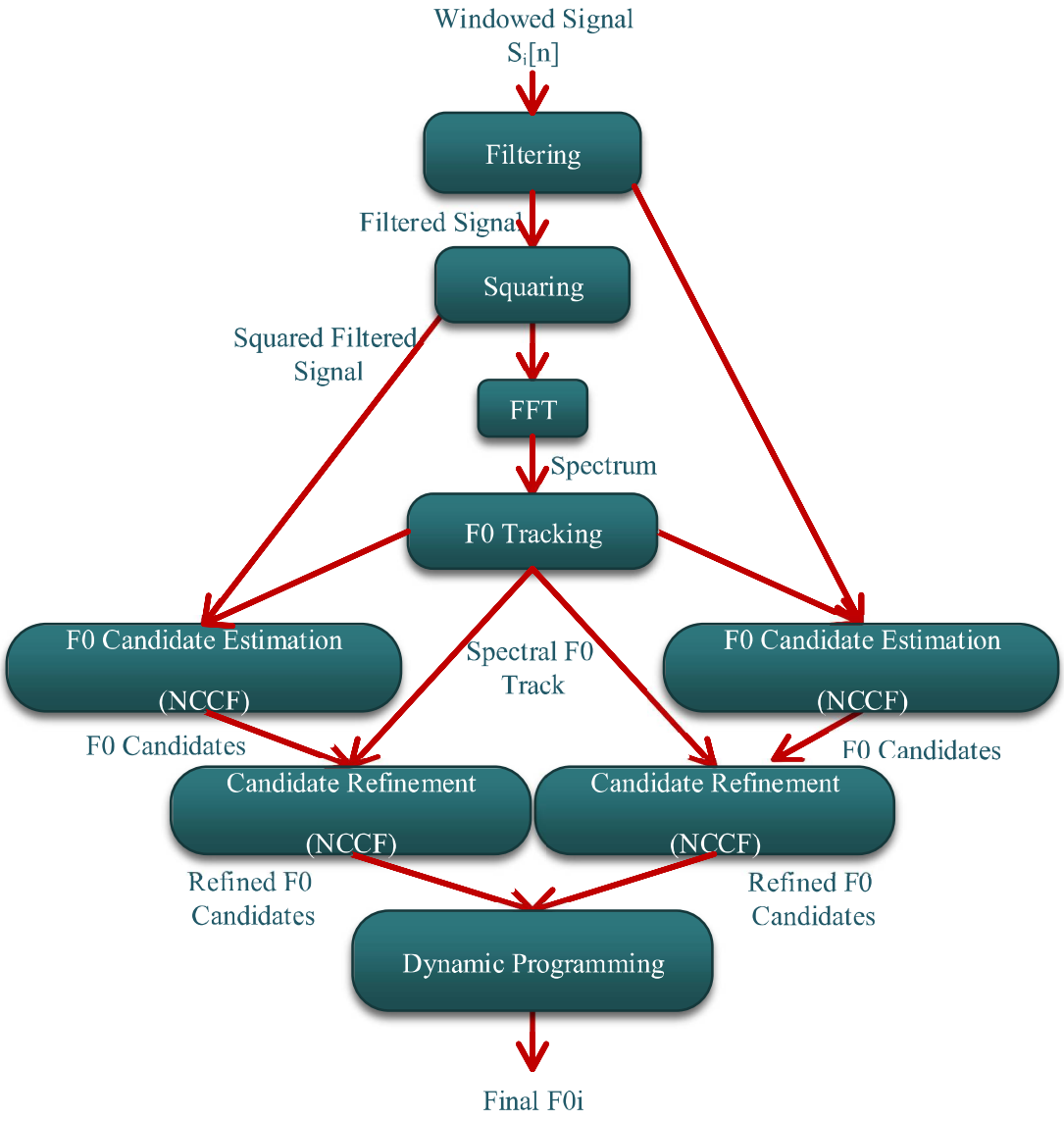

Fig. 2 YAAPT algorithm for pitch detection [48]

Table 1. Note mapping table in Iranian music for fourth octave, as an example

Frequency (Hertz)	Note	Frequency (Hertz)	Note
329.63	Mi	466.16	Si Bemol
339	Fa Koron	480	Si Koron
349.23	Fa	493.88	Si
359	Fa Sori	502	Do Koron
369.99	Sol Bemol	523.25	Do
380	Sol Koron	538	Do Sori
392	Sol	554.37	Re Bemol
403	Sol Sori	570	Re Koron
415.3	La Bemol	587.33	Re
427	La Koron	604	RE Sori
440	La	622.25	Mi Bemol
453	La Sori	640	Mi Koron

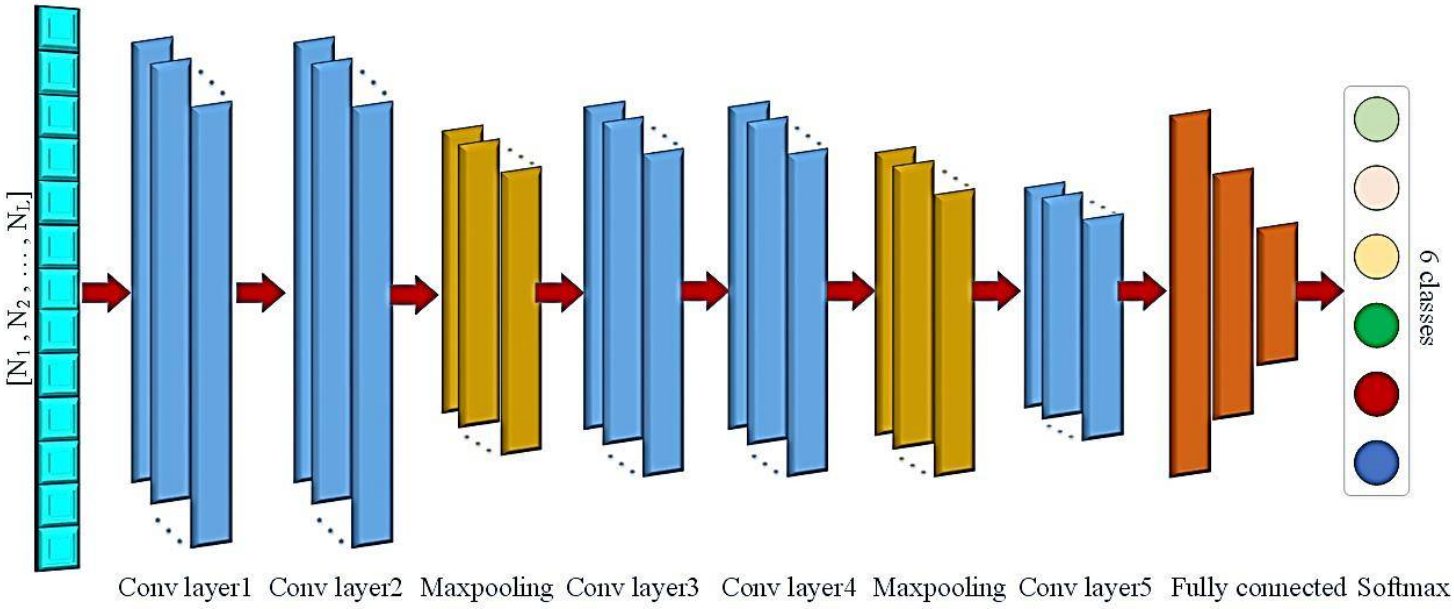

Fig. 3 1DCNN architecture for dastgah recognition

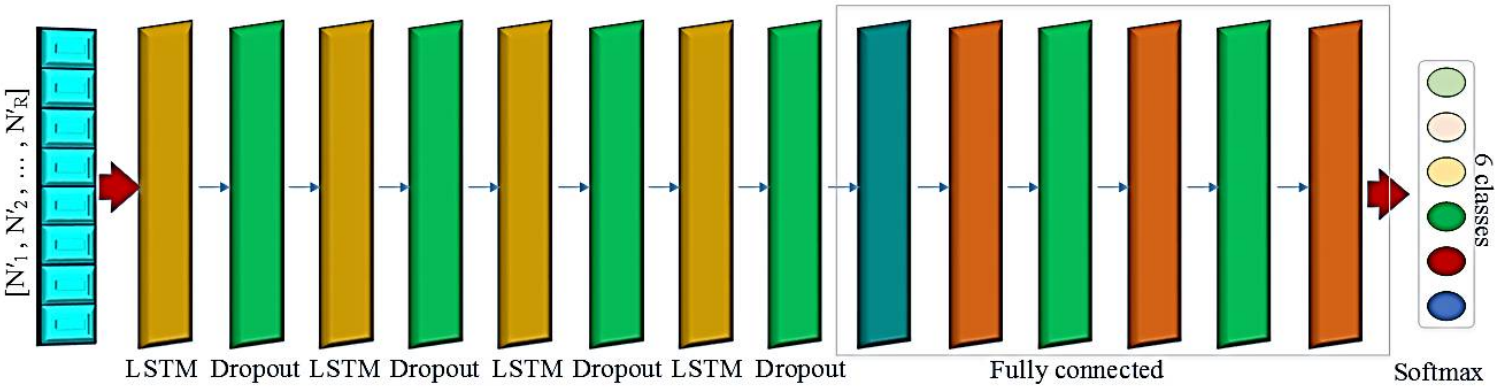

Fig. 4 LSTM architecture for recognizing dastgahs

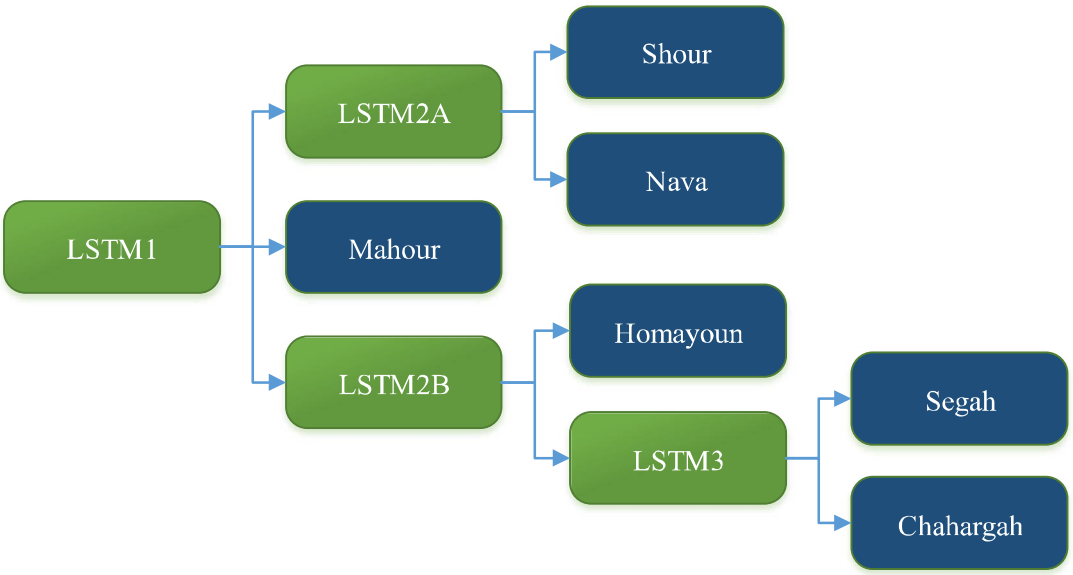

Fig. 5 Hierarchical model of LSTM classifiers

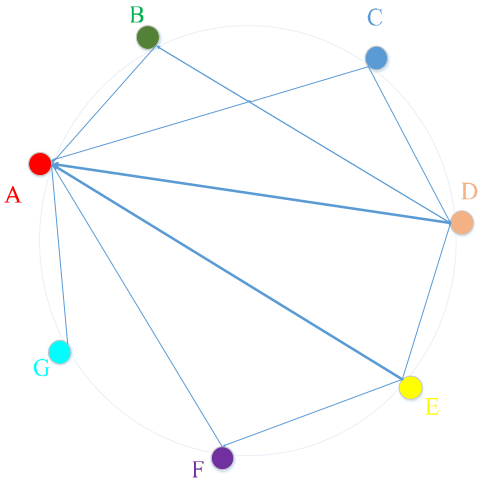

Fig. 6 Graph-based circular representation to display the note sequence

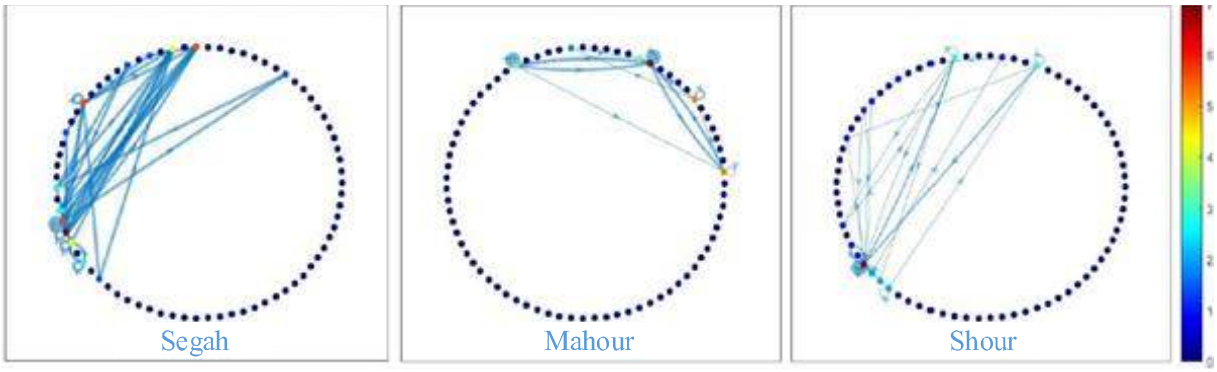

Fig. 7 Graph representation for three examples of Segah, Mahour and Shour

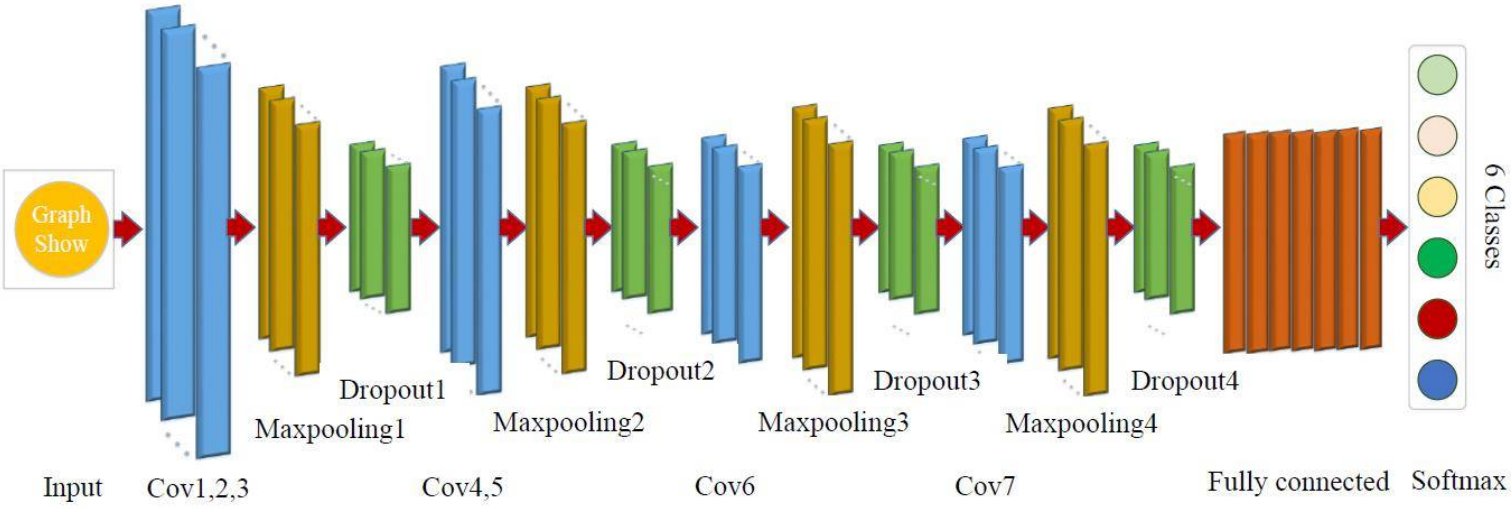

Fig. 8 CNN architecture for Iranian dastgah recognition

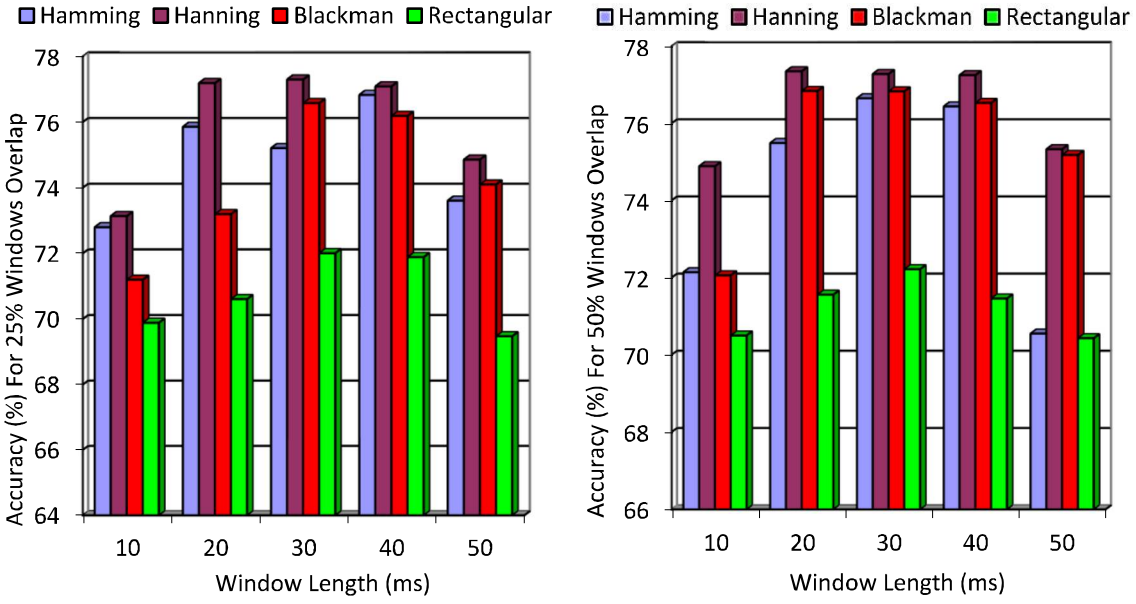

Fig. 9 Classification accuracy for different window lengths, overlaps and window types

Table 2. Accuracy of pitch detection algorithms

Algorithm	Accuracy (%)
YAAPT	77.35
SPICE	76.91
RAPT	76.66
YIN	75.83

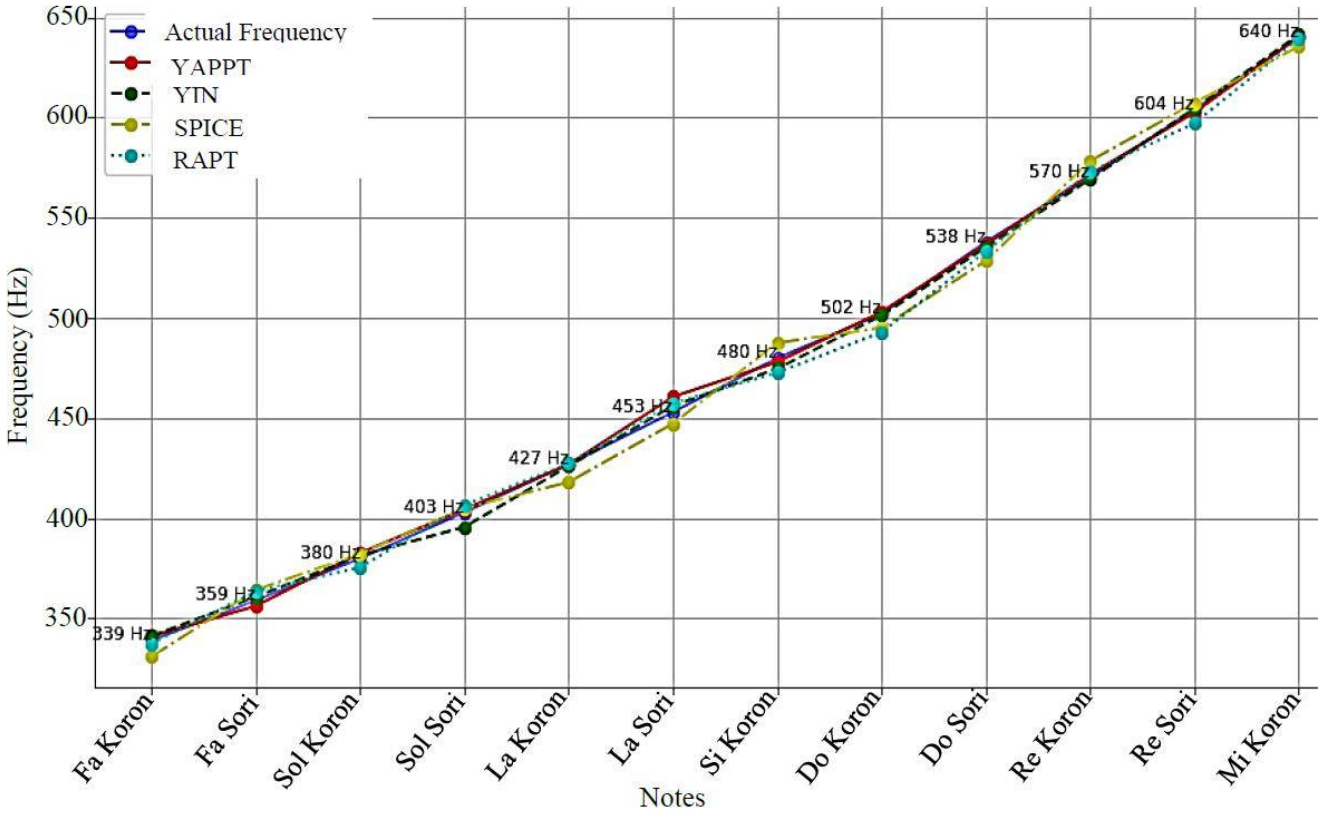

Fig. 10 Comparison of performance for recognizing Quarter-tone Notes

Table 3. Accuracy of hierarchical LSTM structure

Classifiers	Accuracy (%)
LSTM1	84.3
LSTM2A	79.53
LSTM2B	76.21
LSTM3	69.48

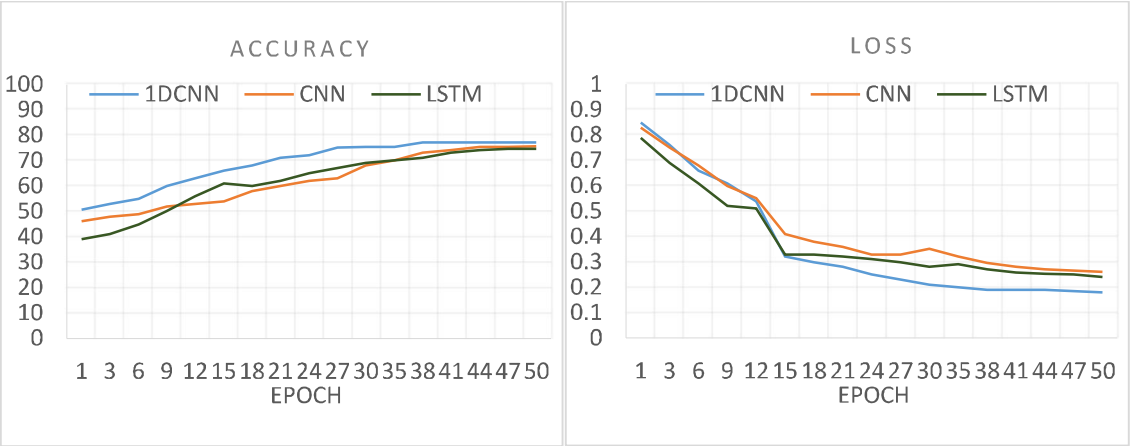

Fig. 11 Accuracy (%) and loss

Table 4. Proposed method results (%)

		F1 Score	Accuracy	STD
Base Model	CNN	72.84	72.54	1.93
	LSTM	74.23	74.58	2.10
	1DCNN	73.47	73.33	1.89
Ensemble Learning method	Stacking	77.39	77.35	1.64
	Bagging	76.65	76.75	1.86
	AdaBoost	76.79	76.89	1.88
	XGBoost	77.04	77.12	1.54
	Majority Voting	76.09	76.20	1.96

Table 5. Proposed method results with data augmentation

	F1 Score (%)	Accuracy (%)	STD
CNN	76.80	76.56	1.49
LSTM	77.87	77.78	1.61
1DCNN	77.25	77.22	1.36
Stacking	80.58	80.44	1.04

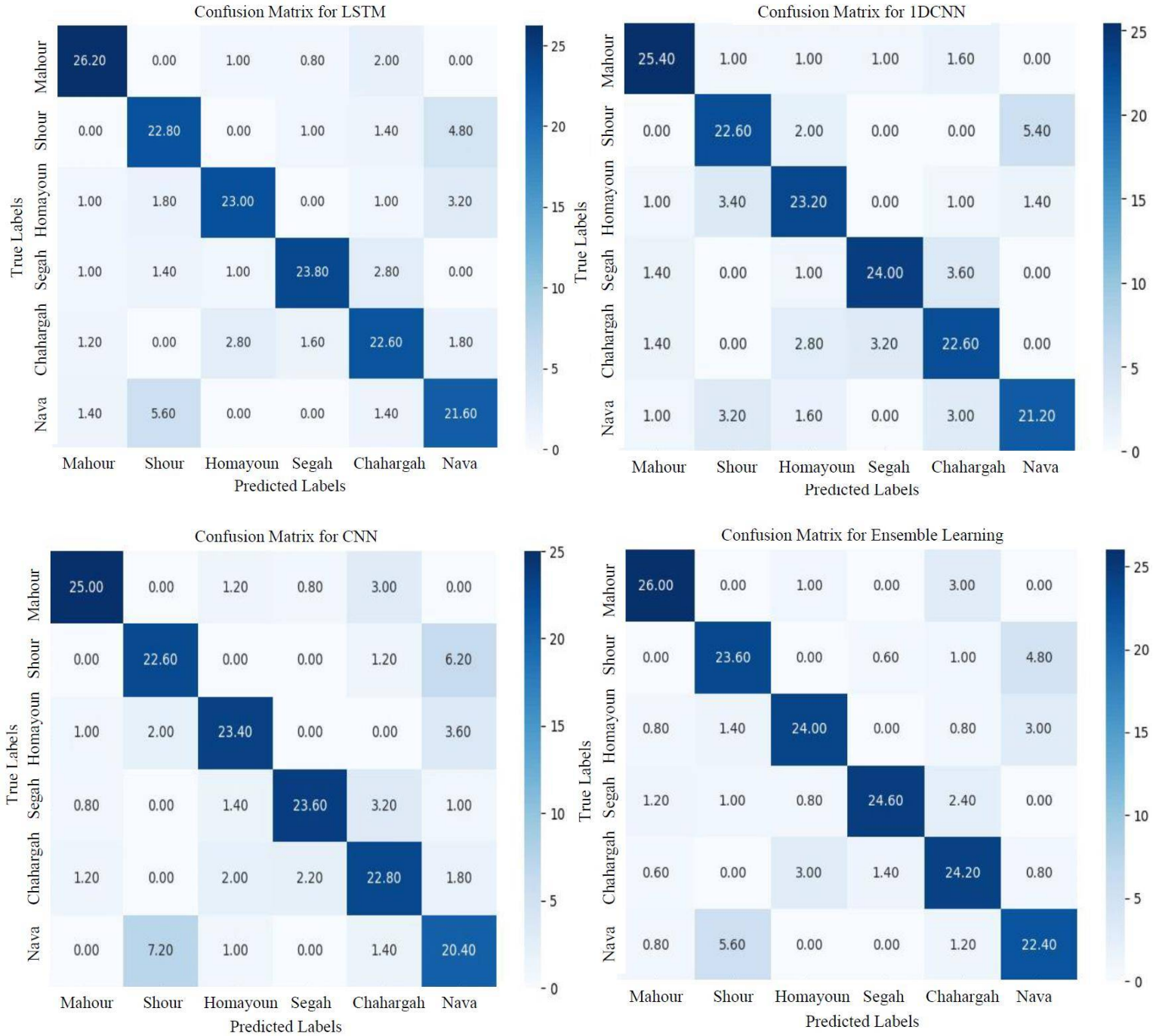

Fig. 12 Confusion matrix for LSTM, 1DCNN, CNN and ensemble learning

Table 6. Comparison of proposed method and other studies

Method	Music/year	Feature	classifier	Accuracy (%)
[38]	Iranian/2022	MFCC	SVM	33.14
[27]	Western/2021	End to end	1DCNN	70.09
[42]	Iranian/2023	End to end	DNN	72.28
[28]	Nigerian/2021	Low-level features	1DCNN	73.65
[43]	Iranian/2022	Pitch	LSTM	74.50
[23]	Western/2020	End to end	BBNN	79.83
Proposed method	Iranian	Pitch	Ensemble	80.44